William Stallings
Computer Organization and Architecture
8th Edition

Chapter 4

Memória Cache

Características

- Posição
- Capacidade
- Unidade de Transferência
- Método de Acesso
- Performance
- Tipo Físico
- Características Físicas
- Organização

Posição

- CPU
- Interna
- Externa

Capacidade

- Tamanho
  - Bytes ou Words

Unidade de Transferência

- Interna
  - Normalmente definida pela capacidade do barramento
- Externa
  - Normalmente, um bloco que é bem maior que um word
- Endereçamento
  - Menor posição que pode ser endereçada unicamente
  - Cluster nos S.O. M$

Métodos de Acesso (1)

- Sequencial
  - Inicia na primeira posição da memória e faz a leitura em ordem
  - Tempo de acesso depende da posição do dado
  - Ex.: Fita
- Direto
  - Blocos individuais tem enderçamento único
  - Acesso é dado pulando a patir de endereços próximos, associado a uma busca
  - Tempo de acesso depende da posição atual e da posição anterior
  - Ex.: Disco rígido (HDD, SSD)

Métodos de Acesso (2)

- Aleatório (Random)
  - Endereços individuais identificam a posição exatamente
  - Tempo de acesso é independente da posição ou de acesso anterior
  - Ex.: RAM
- Associativo
  - Dado é localizado por comparação de conteúdo em um endereço da memória
  - Tempo de acesso é independente da posição ou de acesso anterior
  - Ex.: cache

Hierarquia de Memória

- Registradores
  - CPU
- Interna or Memória Principal
  - Pode incluir um ou mais níveis de cache
  - "RAM"
- Memória Externa
  - Armazenamento persistente

Performance

- Tempo de Acesso - Latência
  - Tempo entre soliciatr o dado apresentado o endereço e receber o dado
- Tempo de Ciclo de Memória
  - Um tempo de recuperação pode ser necessário até a próxima solicitação
  - Tempo de Ciclo =

    Tempo de Acesso + Recuperação
- Taxa de Transferência
  - Taxa na qual o dado pode ser movido

Tipos Físicos

- Semicondutor
  - RAM
  - SSD
- Magnético
  - HDD
  - Fita
  - Disco Flexível
- Ótico
  - CD
  - DVD

Características Fisicas

- Decaimento
- Volatilidade
- Capcaidade de Apagar
- Consumo de Energia

Organização

- Arranjo dos bits em bytes
- Nem sempre óbvio
- Ex.: Interpolado

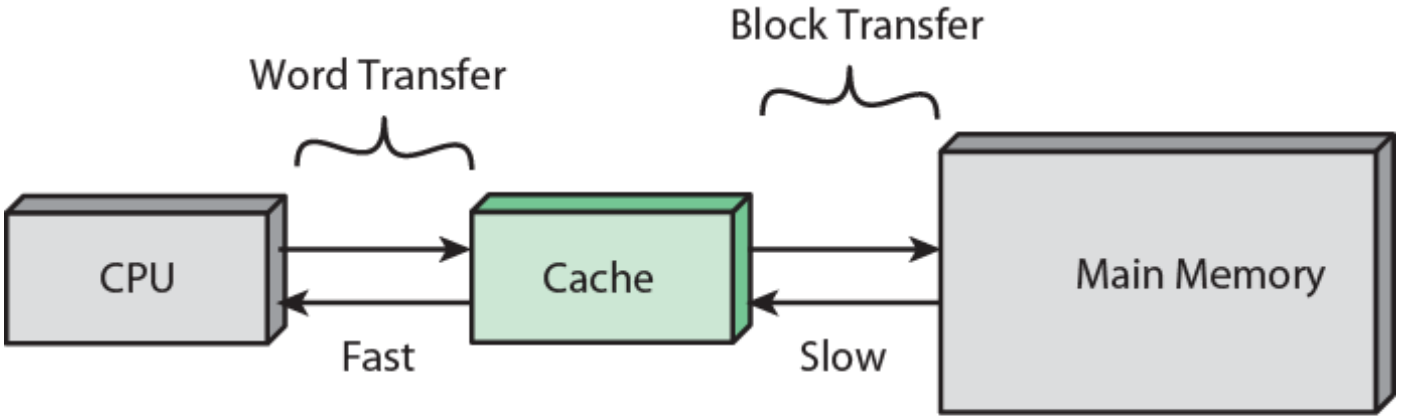
Escolhas

- Quanto?
  - Capacidade
- Quão Rápido?
  - Tempo é dinheiro
- Quão caro?

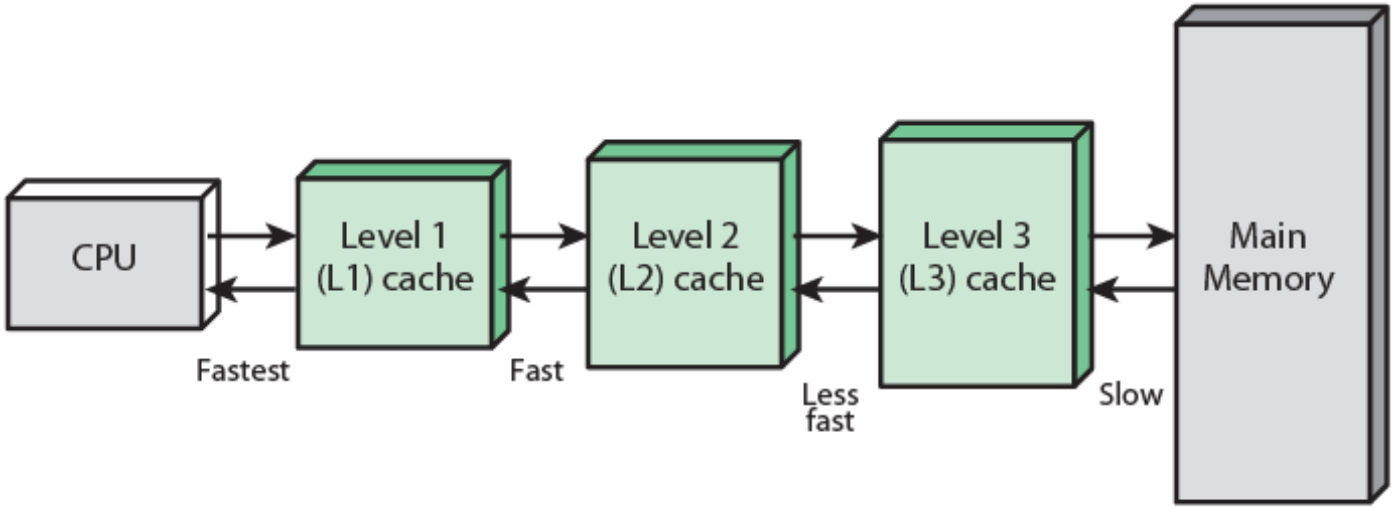
Cache

- Pequena quantidade de memória rápida
- Localizada entre a memória principal e a CPU
- Pode ser um modulo ou estar na CPU

Cache e Memória Principal



(a) Single cache



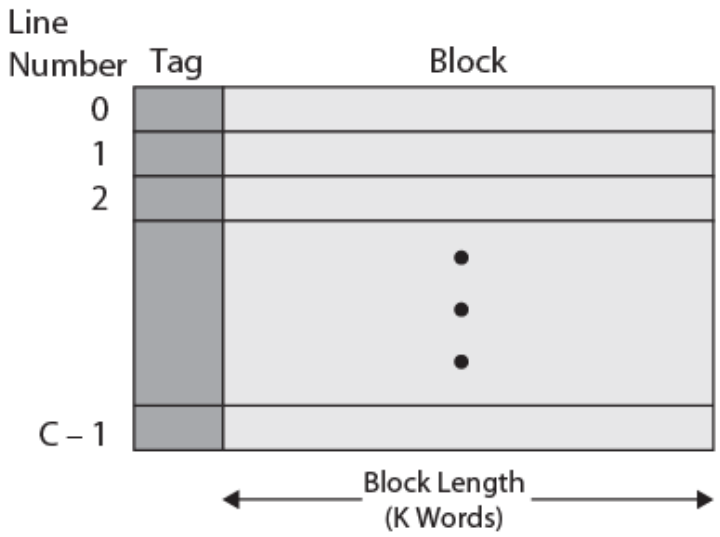
(b) Three-level cache organization

Cache

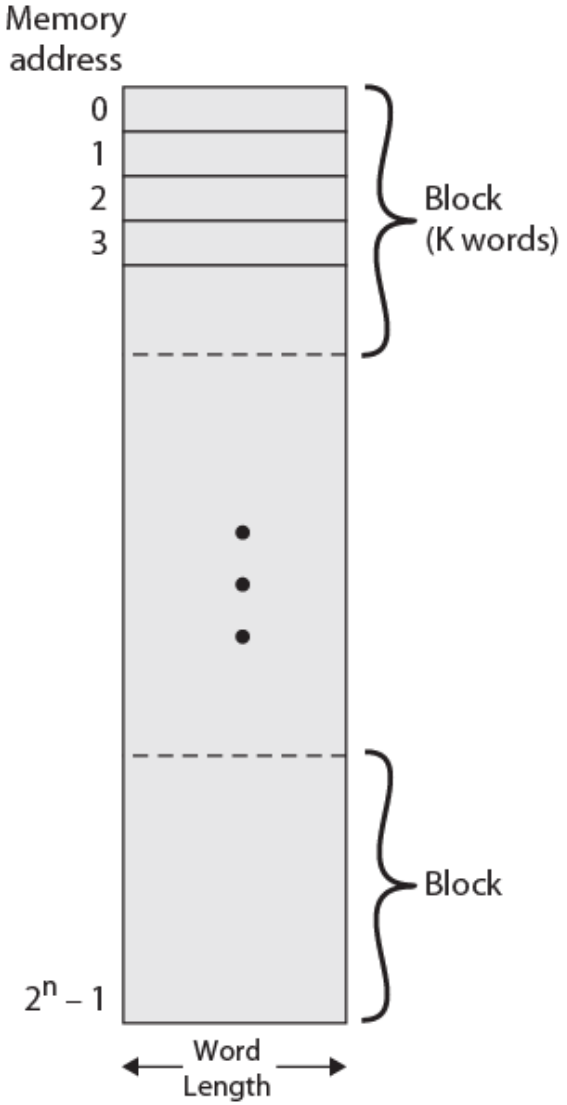
UMC
UM8881F
9642-EYT
M92201

SMART
IN332Q4AO84XG15
061995-T2
Made in USA

Estrutura de Cache/Memory Principal



(a) Cache

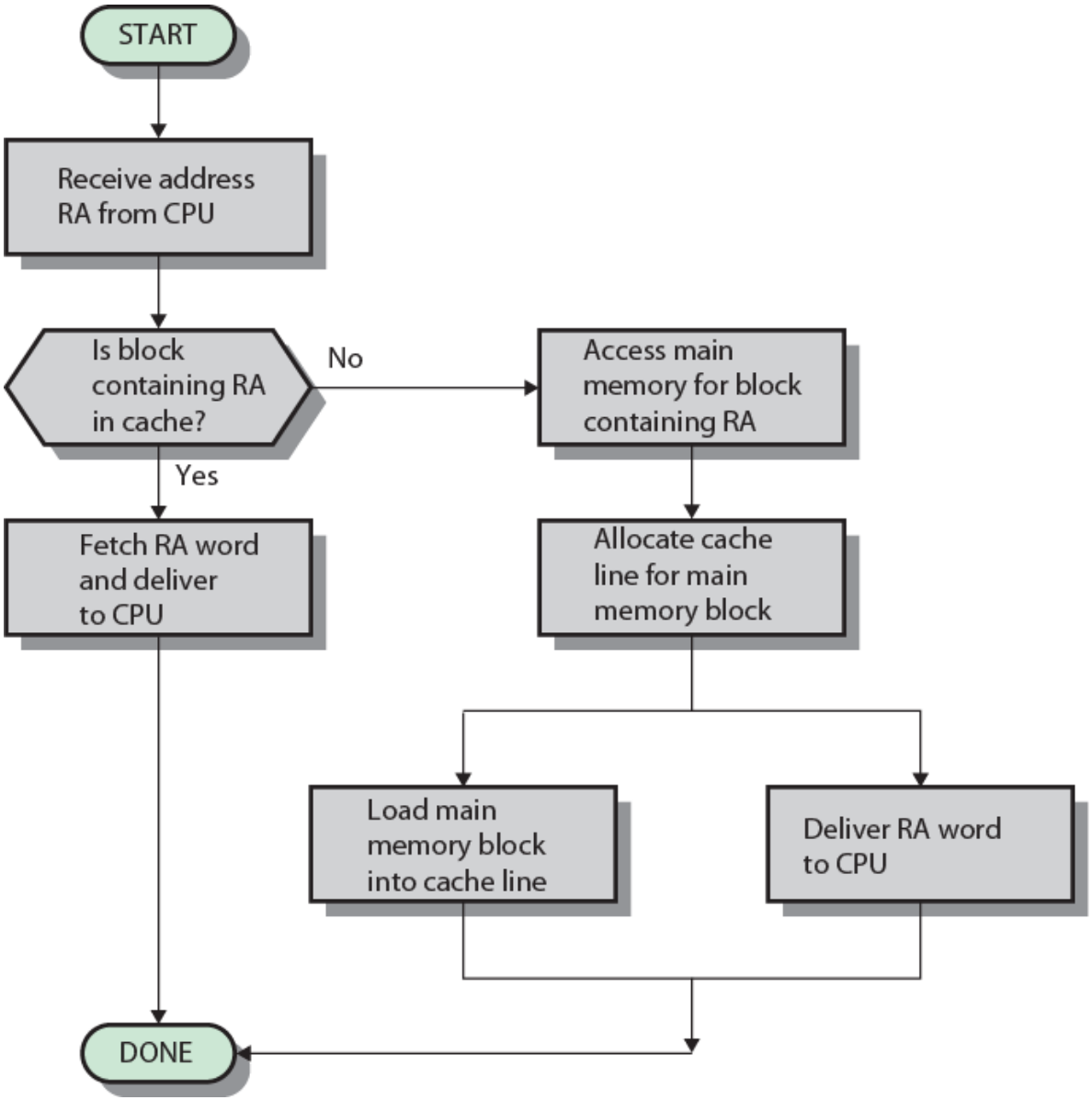


(b) Main memory

Operação de Cache

- CPU solicita conteúdo da memória
- Verifica na cache
- Se está presente, busca da cache (rápido)
- Se não está presente, faz a leitura do dado da memória principal, levando-o à cache
- Então envia o dado da cache para CPU
- Cache inclui tags que identificam qual endereço da memória principal está associado à cache

Operação da Cache - Fluxograma


START
Receive address RA from CPU
Is block containing RA in cache?
No
Yes
Access main memory for block containing RA
Fetch RA word and deliver to CPU
Allocate cache line for main memory block
Load main memory block into cache line
Deliver RA word to CPU
DONE

Cache Design

- Endereçamento
- Tamanho
- Mapeamento
- Algoritmo de Substituição
- Políticas de Escrita
- Tamanho do Bloco
- Cache Multinível

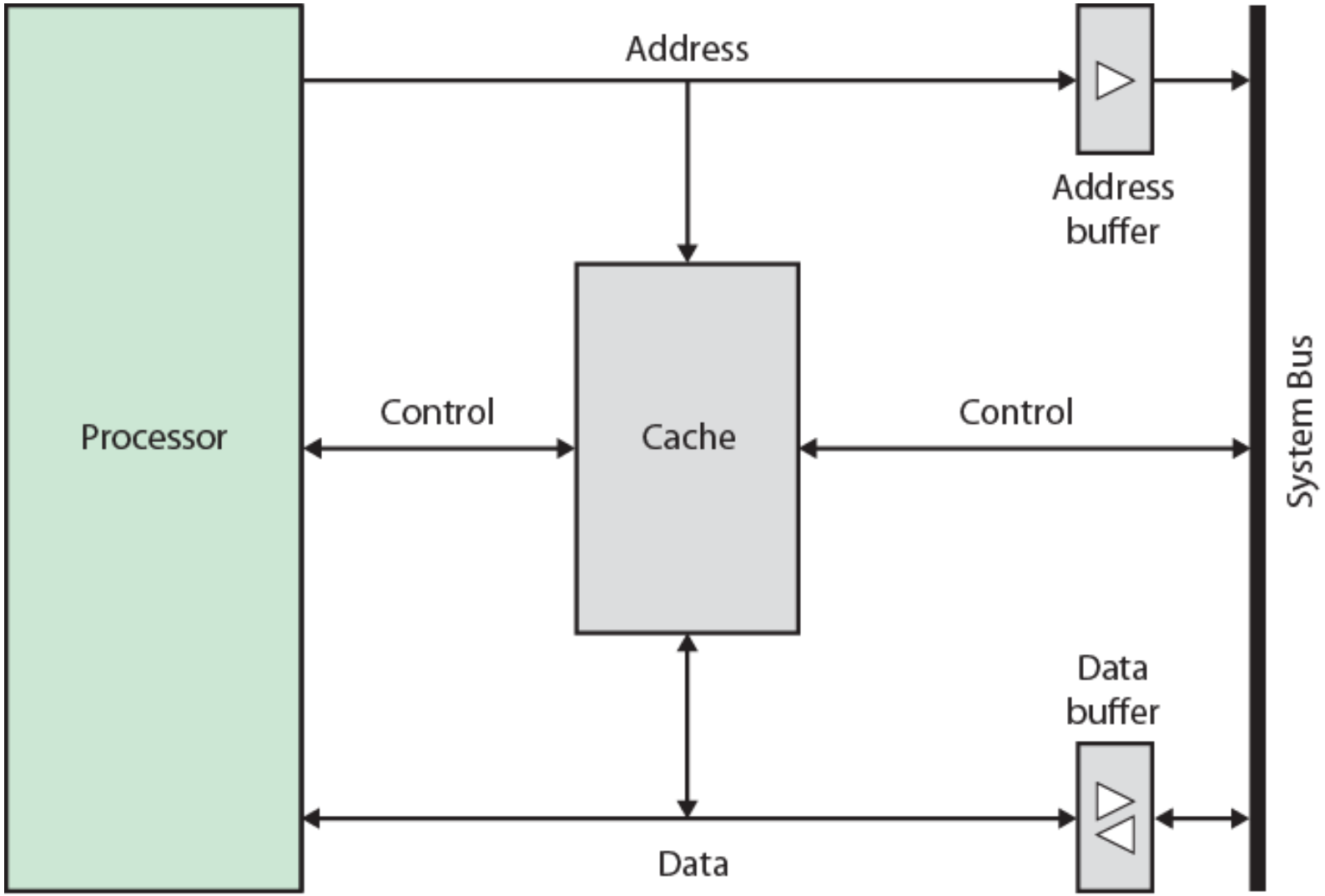
Endereçamento da Cache

- Onde fica a memória cache ?
  - Entre a unidade de processamento e a unidade de gerenciamento de memória virtual (Memory Management Unit – MMU)
  - Entre MMU e memória principal
- Cache armazena dados eferenciando endereçamento da memória principal

Tamanho

- Custo
  - Muita Cache é caro
- Desempenho
  - Muita cache é rápido (até um certo ponto)
  - Buscar dados na cache demandam tempo

Organização Típica de Cache



Disponibilidade Histórica de Cache

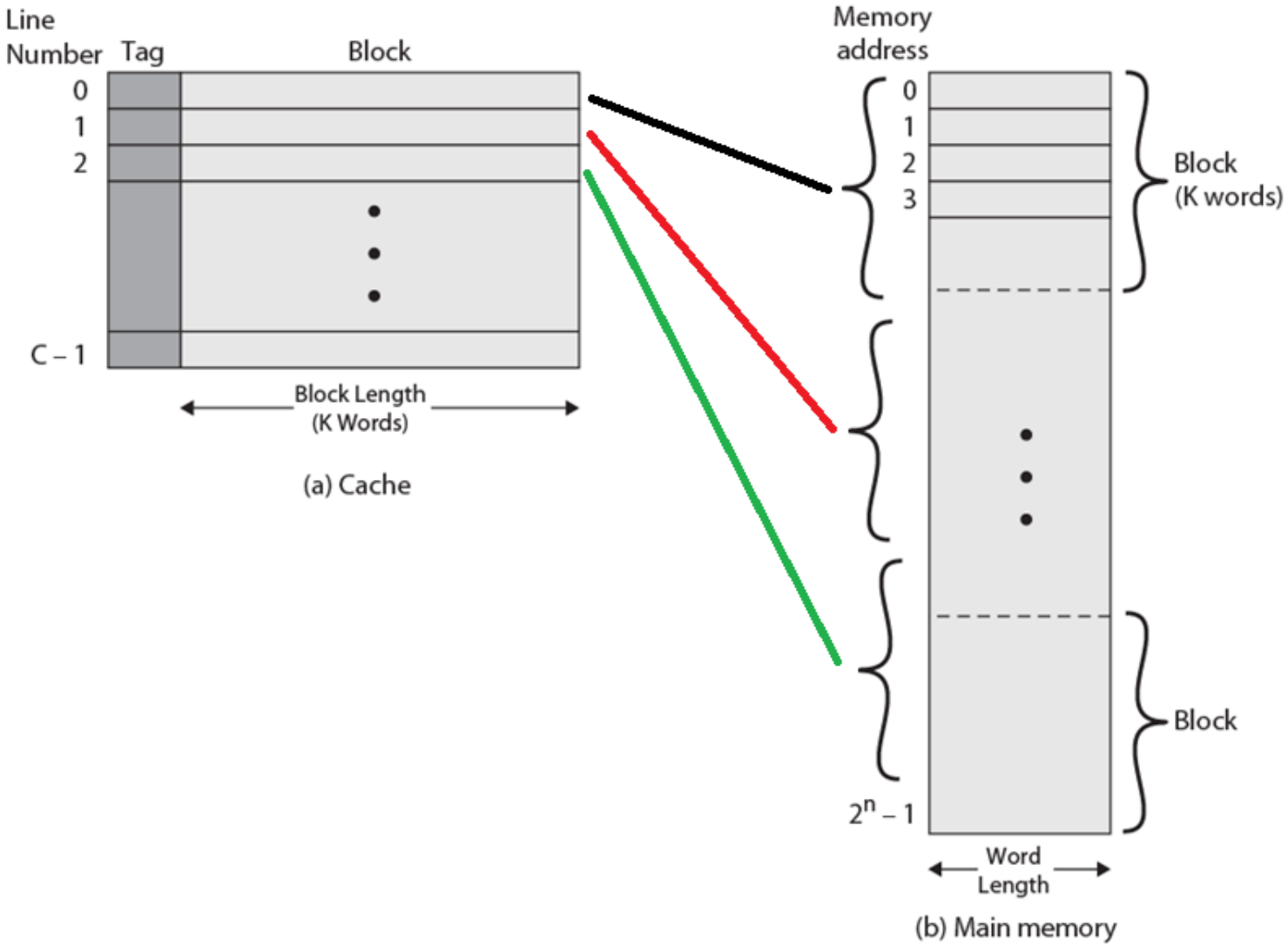
Processador	Tipo de Computador	Ano de Introdução	L1 cache	L2 cache	L3 cache
IBM 360/85	Mainframe	1968	16 to 32 KB	—	—
PDP-11/70	Minicomputer	1975	1 KB	—	—
VAX 11/780	Minicomputer	1978	16 KB	—	—
IBM 3033	Mainframe	1978	64 KB	—	—
IBM 3090	Mainframe	1985	128 to 256 KB	—	—
Intel 80486	PC	1989	8 KB	—	—
Pentium	PC	1993	8 KB/8 KB	256 to 512 KB	—
PowerPC 601	PC	1993	32 KB	—	—
PowerPC 620	PC	1996	32 KB/32 KB	—	—
PowerPC G4	PC/server	1999	32 KB/32 KB	256 KB to 1 MB	2 MB
IBM S/390 G4	Mainframe	1997	32 KB	256 KB	2 MB
IBM S/390 G6	Mainframe	1999	256 KB	8 MB	—
Pentium 4	PC/server	2000	8 KB/8 KB	256 KB	—
IBM SP	High-end server/ supercomputer	2000	64 KB/32 KB	8 MB	—
CRAY MTA[b]	Supercomputer	2000	8 KB	2 MB	—
Itanium	PC/server	2001	16 KB/16 KB	96 KB	4 MB
SGI Origin 2001	High-end server	2001	32 KB/32 KB	4 MB	—
Itanium 2	PC/server	2002	32 KB	256 KB	6 MB
IBM POWER5	High-end server	2003	64 KB	1.9 MB	36 MB
CRAY XD-1	Supercomputer	2004	64 KB/64 KB	1MB	—

Mapeamento Direto

- Um bloco da memória principal é carregado em uma linha especídfica da cache
- Uma linha é examinada para a busca do dados
- Busca na Cache torna-se rápida
- Pode haver subutilização da Cache

Mapeamento Direto da Cache para a Memória Principal

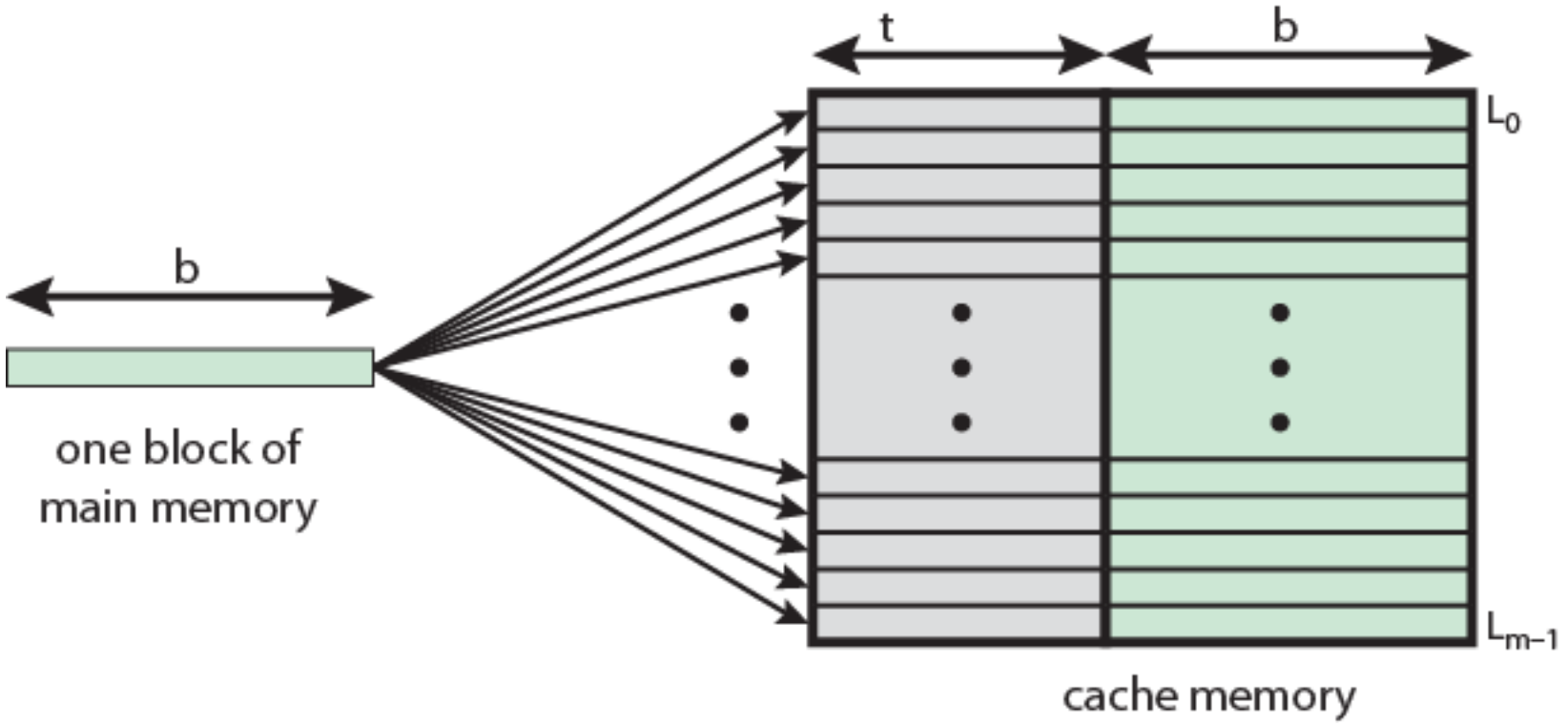
Direct



Mapeamento Associativo

- Um bloco da memória principal pode ser carregado em qualquer linha da cache
- Tag apenas identifica o bloco da memória
- Todas as linhas da Cache são examinadas na busca
- A busca se torna de baixo desempenho
- A utilização da Cache é total

Mapeamento Associativo da Cache para a Memória Principal

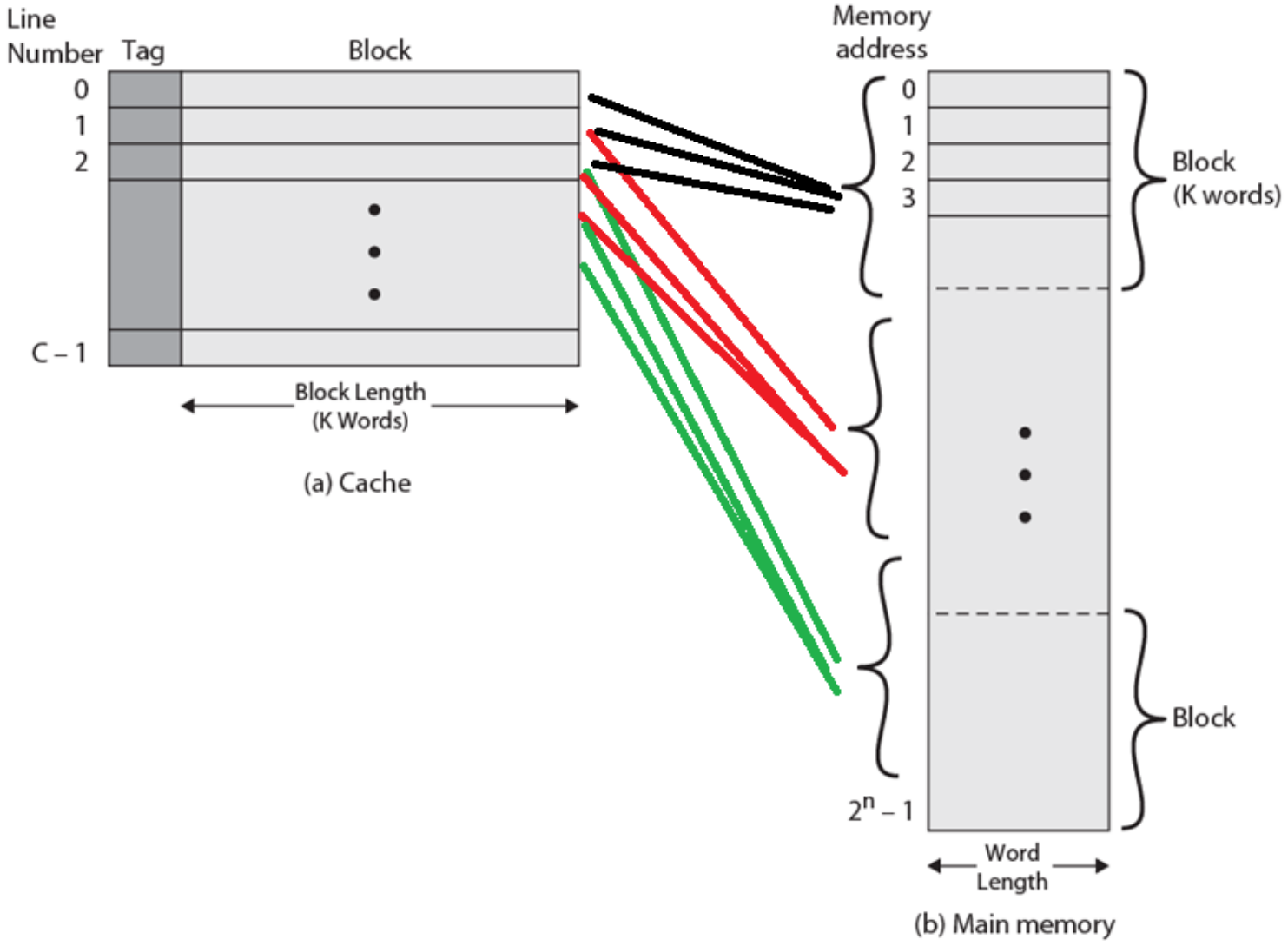


Mapeamento Set Associative

- Um bloco da memória principal pode ser carregado em algumas linhas da cache
- Não é o melhor desempenho possível
- Não é a melhor utilização possível
- Utilização otimizada da cache

Mapeamento Set Associative da Cache para a Memória Principal

Set Associative



Algoritmos de Substituição (1)
Mapeamento Direto

- Sem escolha
- Cada bloco só estará em uma linha específica
- Substitui-se na linha

Algoritmos de Substituição (2) Mapeamento Associativo e Set Associative

- Algoritmo implementado diretamente no hardware
- Random – Aleatório
- First in first out (FIFO) – Primeiro que entra é o primeiro que sai
- Least Recently used (LRU) – Menos Recentemente Usado
  - Substitui o bloco presente na cache há mais tempo
- Least frequently used (LFU) – Menos frequentemente usado
  - Substitui o bloco com menos requisições

Políticas de Escrita

- Ao atulizar o dado na cache, este deve ser atualizado, imediatamente, na memória principal
- Multicore podem ter caches independentes
- E/S podem acessar a memória diretamente

Caches Multinível

- Podem estar todos na CPU ou alguns níveis na CPU e outros Fora
  - L1 no chip, L2 fora
  - L1 e L2 no chip, L3 for a
  - L1, L2 e L3 no chip

Unified v Split Caches

- One cache for data and instructions or two, one for data and one for instructions
- Advantages of unified cache
  - Higher hit rate
    - Balances load of instruction and data fetch
    - Only one cache to design & implement
- Advantages of split cache
  - Eliminates cache contention between instruction fetch/decode unit and execution unit
    - Important in pipelining

Cache Compartilhado

- Intel Smart Cache

Pentium 4 Cache

- 80386 – sem cache no chip
- 80486 – 8k set associative
- Pentium (todas as versões) – L1 cache dividido
  - Dados e Instruções
- Pentium III – L3 cache possível fora da CPU
- Pentium 4
  - L1 cache
    - 8k bytes
    - set associative
  - L2 cache
    - Alimentando ambos L1 caches
    - set associative
  - L3 cache on chip

Pentium 4 Diagrama